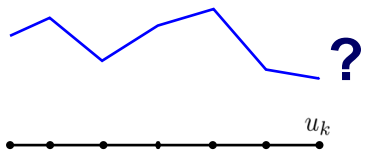
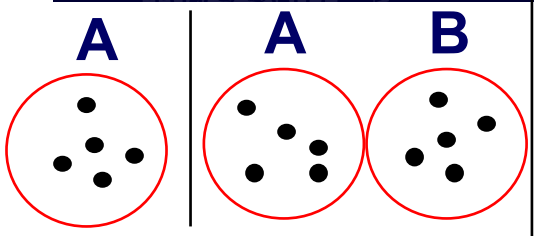
Statistica Descrittiva ed Inferenziale

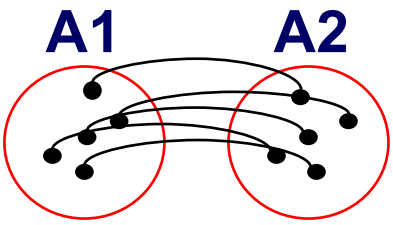
Why Statistics?



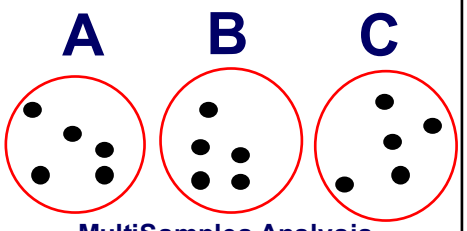
Description and Prediction



Samples Analysis



Paired Samples Analysis



MultiSamples Analysis

Why Statistics?

Formal definition of Probability

A probability measure P on the countable sample space Ω is a set function

$$P : \mathcal{F} \rightarrow [0, 1],$$

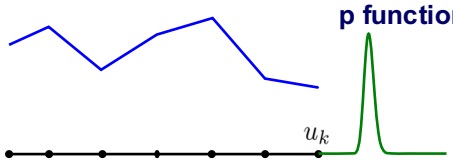
σ-field

satisfying the following conditions

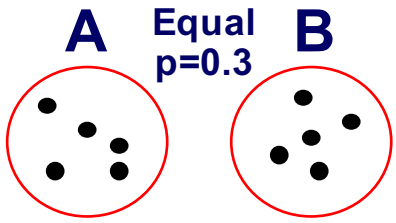
- $P(\Omega) = 1$.
- $P(\omega_i) = p_i$.
- *If* $A_1, A_2, A_3, \ldots \in \mathcal{F}$ *are mutually disjoint, then*

$$P\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} P(A_i).$$

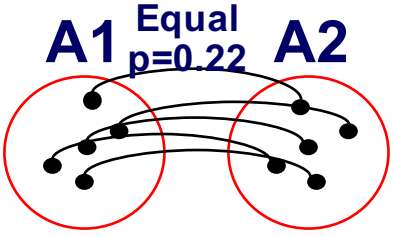
Why Statistics?



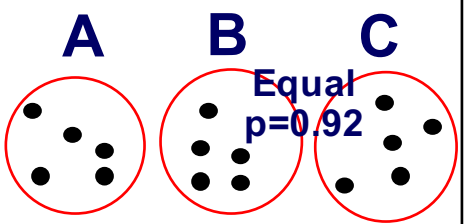
Description and Prediction



Samples Analysis



Samples Analysis



MultiSamples Analysis

Introduzione (1)

Di cosa si occupa la statistica?

Oggetto della statistica sono ***fenomeni collettivi*** che presentano carattere di ***variabilità***

Per fenomeno collettivo si intende un fenomeno che riguarda una grande collezione di elementi = **POPOLAZIONE**

Elementi della popolazione = **UNITÀ STATISTICHE**

Introduzione (2)

La popolazione è troppo vasta per poter essere studiata nella sua globalità → dalla popolazione viene estratto un **CAMPIONE** di n elementi

Sul campione vengono rilevate/misurate alcune caratteristiche. I risultati di questa misura costituiscono i **DATI**

La statistica permette di trarre conclusioni sull'intera popolazione a partire dai dati ottenuti sul campione

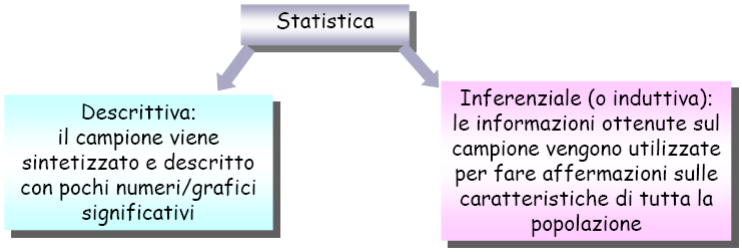
A causa della casistica ridotta non possiamo essere certi delle nostre conclusioni→ si specifica il grado di certezza in termini di probabilità

Statistica Descrittiva e Statistica Inferenziale (1)

Scopo principale della statistica consiste nel compiere un'inferenza circa l'intera popolazione a partire dal campione

Per fare questo, per prima cosa, bisogna descrivere e sintetizzare i dati, con pochi numeri o grafici significativi

Si distinguono due grandi rami della statistica:



Statistica Descrittiva e Statistica Inferenziale (2)

Nelle applicazioni, Statistica Descrittiva e Statistica Inferenziale non possono essere completamente separate.

Infatti i problemi di inferenza statistica vengono affrontati secondo il seguente schema:

Statistica Descrittiva di un CAMPIONE (SAMPLE) di dati

Statistica Descrittiva

Scopo: descrivere il campione (dati) in modo sintetico ed efficace mediante tabelle, grafici, numeri

Premessa: Le caratteristiche che osserviamo sul campione variano da un'unità di osservazione all'altra → variabili

Le variabili possono essere **discrete** o **continue**

Variabili discrete: assumono un numero finito o un'infinità numerabile di valori

Variabili continue: possono assumere qualsiasi valore

Tabelle e Grafici di Frequenza (1)

Un primo utile sistema per riassumere i dati è la costruzione di tabelle e grafici di frequenza

Esempio nel discreto: lancio di un dado

Il risultato del lancio è una variabile discreta (può assumere uno dei seguenti valori 1 2 3 4 5 6)

50 lanci → ottengo una sequenza di 50 (n) numeri

Sintetizzo i dati costruendo una tabella

Tabelle e Grafici di Frequenza (2)

Esempio nel discreto *(continua)*

risultato del lancio	frequenza (f)	frequenza relativa (f/n)
1	9	0.18
2	12	0.24
3	6	0.12
4	8	0.16
5	10	0.20
6	5	0.10
	$\sum f = n$	$\sum (f/n) = 1.00$

prima colonna: possibili risultati del lancio

seconda colonna: numero totale di volte in cui è stato ottenuto quel risultato (**frequenza assoluta**)

terza colonna: frequenza assoluta del risultato divisa per il numero totale (n) di osservazioni (**frequenza relativa**)

Tabelle e Grafici di Frequenza (3)

Esempio nel discreto *(continua)*

risultato del lancio	frequenza (f)	frequenza relativa (f/n)
1	9	0.18
2	12	0.24
3	6	0.12
4	8	0.16
5	10	0.20
6	5	0.10

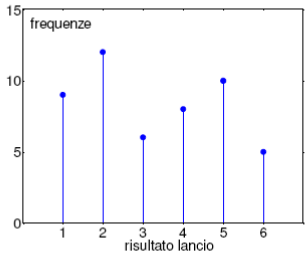
seconda colonna: **distribuzione di frequenze**

terza colonna: **distribuzione di frequenze relative**

Tabelle e Grafici di Frequenza (4)

Esempio nel discreto *(continua)*

La distribuzione di frequenza e la distribuzione di frequenza relativa possono essere rappresentate graficamente mediante **diagrammi a bastoncino**



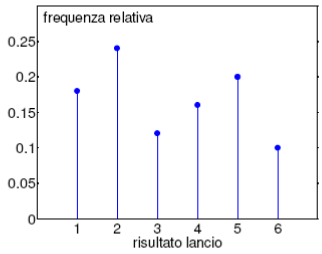


Tabelle e Grafici di Frequenza (5)

Esempio nel continuo

Campione di 200 uomini (n=200) estratto da una certa popolazione. Rileviamo l'altezza in cm.

La variabile osservata è continua. Non ha senso parlare di frequenza del singolo valore poiché non c'è alcuna possibilità di osservare due stature esattamente uguali.

L'intervallo che contiene tutti i valori osservati viene suddiviso in un certo numero di sottointervalli (classi) e si contano quante osservazioni cadono nei diversi sottointervalli.

Tabelle e Grafici di Frequenza (6)

Esempio nel continuo (continua)

Limiti intervallo (cm)	Valore centrale (cm)	frequenza (f)	frequenza relativa (f/n)
141.5-148.5	145	2	0.01
148.5-155.5	152	7	0.035
155.5-162.5	159	22	0.11
162.5-169.5	166	13	0.065
169.5-176.5	173	44	0.22
176.5-183.5	180	36	0.18
183.5-190.5	187	32	0.16
190.5-197.5	194	13	0.065
197.5-204.5	201	21	0.105
204.5-211.5	208	10	0.05
		$\sum f = n$	$\sum (f/n) = 1.00$

Tabelle e Grafici di Frequenza (7)

Quante classi (sottointervalli) vi devono essere?

- compromesso ragionevole tra una distribuzione troppo dettagliata ed una troppo sintetica

Le classi vengono in genere scelte in modo che il valore centrale sia un numero intero

In quale classe viene posizionata una osservazione che cade al limite tra due classi?

In genere la si pone nella classe superiore

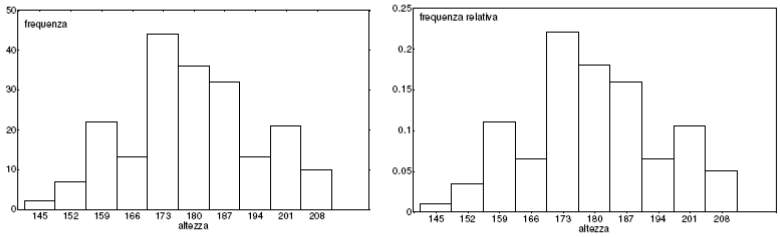
Ad esempio il valore 162.5 viene posto nella classe 162.5-169.5.

Si tratta quindi di sottointervalli del tipo [ )

Tabelle e Grafici di Frequenza (8)

Esempio nel continuo (continua)

I dati raggruppati in sottointervalli possono essere rappresentati graficamente mediante **istogrammi**



istogramma della distribuzione di frequenze

base = ampiezza della classe

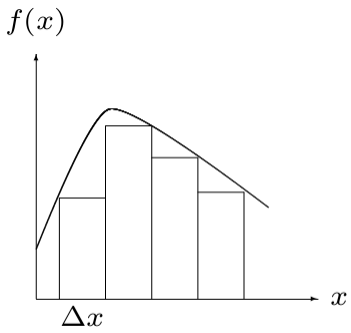
altezza = frequenza assoluta

istogramma della distribuzione di frequenze relative

base = ampiezza della classe

altezza = frequenza relativa

Obiettivo: Descrizione di un Istogramma



Indici di descrizione statistica di un campione

INDICI

- di posizione
  - MODA
  - MEDIANA
  - MEDIA
- di dispersione
  - SCARTO QUADRATICO MEDIO
  - VARIANZA
  - RANGE
- di forma
  - ASIMMETRIA (SKEWNESS)
  - CURTOSI (KURTOSIS)

Misure di posizione : La Media Aritmetica (1)

La media aritmetica (m) è la più comune misura di posizione

Le osservazioni (x_1, x_2,…x_n) vengono sommate tra di loro, quindi la somma divisa per n (cioè per il numero di osservazioni):

$$m = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

Per calcolare l'altezza media del nostro campione di 200 individui dobbiamo sommare le 200 osservazioni e dividere la somma per 200

Centro di una distribuzione

dato un insieme di n elementi $\{x_1, x_2, \ldots x_N\}$

- Si dice **media aritmetica semplice** di N numeri il numero che si ottiene dividendo la loro somma per N.

$$\bar{x} = \frac{x_1 + x_2 + \mathrm{L} \quad + x_N}{N}$$

dato un insieme di m elementi $\{x_1, x_2, \ldots x_m\}$, e

dato un insieme di m di numeri reali $\{p_1, p_2, \ldots p_m\}$

- Si dice **media aritmetica pesata**

$$\bar{x} = \frac{x_1 \cdot p_1 + x_2 \cdot p_2 + \ldots + x_m \cdot p_m}{p_1 + p_2 + \ldots + p_m}$$

che utilizza un peso p_j o la frequenza di ogni dato x_j per j=1,…,m

Misure di posizione : La Media Aritmetica (2)

Se è nota solo la distribuzione di frequenza per sottointervalli (non le singole osservazioni)

→ si calcola una **media approssimata**

Sia f_1 il numero di osservazioni che cadono nel sottointervallo 1;

tali osservazioni vengono approssimate dal valore centrale del sottointervallo (x_1)

analogamente per tutti gli altri sottointervalli

$$m \cong \frac{1}{n} \cdot \sum_{i=1}^{k} f_i x_i = \sum_{i=1}^{k} x_i \left(\frac{f_i}{n} \right)$$

f_i / n: frequenza relativa del sottointervallo *i*-esimo

k : numero dei sottointervalli

Esempio di media pesata

La media della lunghezza di un gruppo di $f_1 = 7$ neonati ⇨ $m_1 = 48.0$ cm
e di altri $f_2 = 3$ neonati ⇨ $m_2 = 49.5$ cm.

Per calcolare la media delle lunghezze dell'insieme totale di **10 neonati** pur senza avere la conoscenza dei valori delle lunghezze individuali, si utilizzano le proprietà della media aritmetica :

la somma delle lunghezze dei primi 7 è	48.0 × 7 = 336.0
la somma delle lunghezze dei secondi 3 è	49.5 × 3 = 148.5
la somma delle lunghezze di tutti i 10 è	= 484 .5

La media della lunghezza di tutti i 10 neonati è = 484.5/10 = 48.45

Ovvero **Media = $(f_1 \times m_1 + f_2 \times m_2)/(f_1 + f_2)$**
Media = $(7 \times 48.0 + 3 \times 49.5)/(7+3)$

esempio di media aritmetica

Lunghezza(cm) in un campione di 60 neonati.

51.0	49.4	49.0	52.5	51.5	51.8
46.5	47.8	49.7	44.5	49.8	53.0
48.7	50.0	52.9	50.8	46.2	48.9
54.5	48.2	48.9	51.2	49.5	56.3
46.0	52.2	47.0	50.8	50.0	52.5
51.2	51.1	54.7	52.3	48.2	50.8
55.0	50.2	50.3	47.7	48.5	53.8
50.2	53.4	47.4	50.5	51.7	49.5
44.4	49.2	50.5	49.5	52.9	50.5
54.0	46.5	51.5	50.9	51.6	52.7

la media aritmetica dei primi 6 valori di lunghezza di 6 neonati è:
$\bar{x}$ = (51.0+49.4+49.0+52.5+51.5+51.8)/6 = 305.2/6 = 50.87
la media aritmetica di tutti i 60 valori di lunghezza è:
= (55.9+51.3+53.0+50.5+54.9+53.4+…+53.8)/60 = 3021.8 /60
$\bar{x}$ = 50.363

La media aritmetica di N dati distinti è …

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$$

MEDIA per dati raggruppati in classi

ALTEZZA(cm) di un campione di 60 neonati.

limiti di classe	x_i	$f(x_j)$	$x_i f(x_j)$
44.25- 45.75	45.0	2	90.0
45.75- 47.25	46.5	5	232.5
47.25- 48.75	48.0	7	336.0
48.75- 50.25	49.5	14	693.0
50.25- 51.75	51.0	16	816.0
51.75- 53.25	52.5	9	472.5
53.25- 54.75	54.0	5	270.0
54.75- 56.25	55.5	1	55.5
56.25- 57.75	57.0	1	57.0
		60	3022.5

Nell'*esempio* del campione di 60 misure di lunghezza dei neonati:

$$\bar{x} = \frac{45.0 \times 2 + 46.5 \times 5 + \ldots 57.0 \times 1}{2 + 3 + \ldots 1} = \frac{3022.5}{60} = 50.375$$

La media per dati raggruppati in m classi è ...

$$\bar{x} = \frac{\sum_{j=1}^{m} x_j \cdot f(x_j)}{\sum_{j=1}^{m} f(x_j)}$$

dove **m** è il numero di classi e ,

$\sum_{j=1}^{m} f(x_j) = N$ se $f(x_i)$ indica le frequenze assolute,

oppure $\sum_{j=1}^{m} f(x_j) = 1$ se $f(x_i)$ indica le frequenze relative.

media aritmetica e mediana

Si consideri un campione di valori di VES (*velocità di eritrosedimen-tazione*, mm/ora) misurati in 7 pazienti

{8, 5, 7, 6, 35, 5, 4}

In questo caso, **la media** ($\bar{x}$ = 10 mm/ora) **non è** un valore **tipico** della distribuzione: soltanto un valore su 7 è superiore alla media!

Conviene usare come indice del centro **la mediana**, definita come quel valore che divide a metà la distribuzione, sicché **l'insieme dei valori è per metà minore e per metà maggiore della mediana**.

Per **calcolare la mediana** si dispongono i dati in ordine crescente:
ordine originale: {8, 5, 7, 6, 35, 5, 4}
ordine crescente: {4, 5, 5, 6, 7, 8, 35}

mediana

Se n è **dispari**, la mediana è il valore che occupa la posizione (n+1)/2 nell'insieme ordinato.

Nell'*esempio*, poiché (n+1)/2=4, la mediana è 6 mm/ora, ed è tipica nel senso che si avvicina a buona parte dei valori del campione.

Se n è **pari**, la mediana è la media dei valori che occupano le posizioni (n/2) ed [(n/2)+1] nell'insieme ordinato dei numeri.

Se, nell'***esempio***, si esclude il valore più alto, si ottiene l'insieme ordinato {4, 5, 5, 6, 7, 8},

(n/2)=3 e [(n/2)+1]=4,

e la mediana vale (5+6)/2=5.5.

Frattili di una distribuzione

Una distribuzione può essere descritta per mezzo dei suoi **frattili**.

Si dice frattile (sinonimi: **centile, percentile e quantile**) p-*esimo* di una distribuzione quel valore xp tale che la frequenza relativa cumulata $F(x_p) = p$.

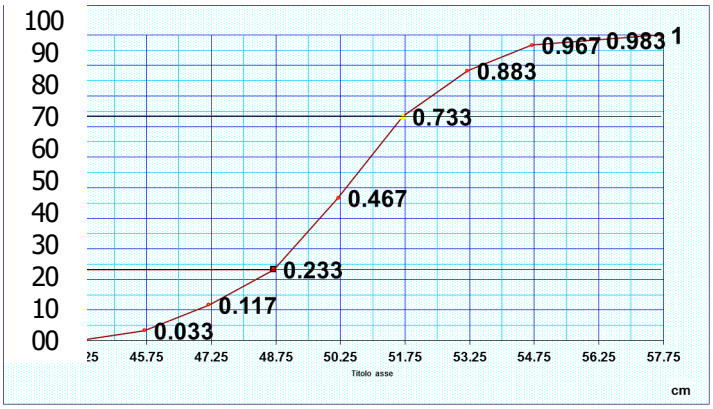
Ad esempio, **il 50° centile** di una distribuzione è il valore che, sull'asse dei numeri reali, ha alla sua sinistra il 50% dei valori della distribuzione, e **coincide con la mediana**.

Il 10° centile è il valore che ha alla sinistra il 10% della distribuzione.

ALTEZZA(cm) di un campione di 60 neonati.

limiti di classe	x_i	$f(x_j)$	$x_i f(x_j)$
44.25- 45.75	45.0	2	90.0
45.75- 47.25	46.5	5	232.5
47.25- 48.75	48.0	7	336.0
48.75- 50.25	49.5	14	693.0
50.25- 51.75	51.0	16	816.0
51.75- 53.25	52.5	9	472.5
53.25- 54.75	54.0	5	270.0
54.75- 56.25	55.5	1	55.5
56.25- 57.75	57.0	1	57.0
		60	3022.5

Nei **grafici cumulati**, i valori riportati sull'asse verticale indicano la **frequenza** delle rilevazioni con **valore pari o minore** ai valori in corrispondenza sull'asse orizzontale



La Moda

Più di rado si incontra una terza misura di posizione, la moda; è il *valore che si verifica più spesso (frequenza assoluta più elevata);* la modalità della variabile in cui si registra il maggior numero di casi.

Quanto sono usualmente lunghi i bimbi alla nascita?

Guardando i dati a nostra disposizione, è subito evidente maggior numero (16) di bimbi è lungo tra i 50.3 cm e i 51.7 cm.

la classe modale è dunque 50.25-51.75.

Se la distribuzione ha più di due valori massimi o se la frequenza più alta riscontrata nell'insieme considerato non supera di molto le altre la moda non è un buon indicatore di tendenza centrale.

La moda

Lunghezza supina (cm) in un campione di 60 neonati.
Valori ottenuti con l'infantometro Harpenden.

Estremi di classe	Valore Centrale	Freq Semplici n	Freq Semplici %	Freq cumulate n	Freq cumulate %
44.3-45.7	45.0	2	0.033333	2	0.033333
45.8-47.2	46.5	5	0.083333	7	0.116667
47.3-48.7	48.0	7	0.116667	14	0.233333
48.8-50.2	49.5	14	0.233333	28	0.466667
50.3-51.7	51.0	16	0.266667	44	0.733333
51.8-53.2	52.5	9	0.15	53	0.883333
53.3-54.7	54.0	5	0.083333	58	0.966667
54.8-56.2	55.5	1	0.016667	59	0.983333
56.3-57.7	57.0	1	0.016667	60	1

Nella classe 50.3-51.7 , piu' vicino alla casse con freq=14

$$50{,}25+\frac{1{,}5\times|16-14|}{|16\text{-}14|+|16\text{-}9|}=50{,}583$$

quale misura di posizione usare?

A quale misura di tendenza centrale ci riferiamo?

- Il proprietario di una ditta afferma "Lo stipendio medio nella nostra ditta è 2.700 euro"
- Il sindacato dei lavoratori dice che "lo stipendio mensile è di 1.700 euro".
- L'agente delle tasse dice che "lo stipendio è stato quasi sempre di 2.200 euro". Queste risposte diverse sono state ottenute tutte dai dati della seguente tabella.

Stipendio mensile	N° di lavoratori
1.300	2
1.700	22
2.200	19
2.600	3
6.500	2
9.400	1
23.000	1

Media aritmetica= lire 2.700
Mediana = lire 2.200
Moda = lire 1.700

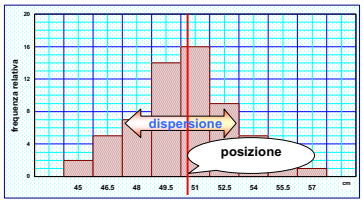
interpretazione delle misure di posizione

- La **media aritmetica** indica che, se il denaro fosse distribuito in modo che ciascuno ricevesse la stessa somma, ciascun dipendente avrebbe avuto 2.700 euro
- La **moda** ci dice che la paga mensile più comune è di 1.700.euro
- La moda si considera spesso come il valore tipico dell'insieme di dati poiché è quello che si presenta più spesso. **Non tiene però conto degli altri valori** e spesso in un insieme di dati vi è **più di un valore** che corrisponde alla definizione di moda.
- La **mediana** indica che circa metà degli addetti percepiscono meno di 2.200.euro, e metà di più.
- La mediana **non è influenzata dai valori estremi** eventualmente presenti ma solo dal fatto che essi siano sotto o sopra il centro dell'insieme dei dati.

Statistica Descrittiva

- Intervallo di variazione
- Devianza
- Varianza
- Deviazione Standard
- Intervallo interquartile

dispersione di una distribuzione

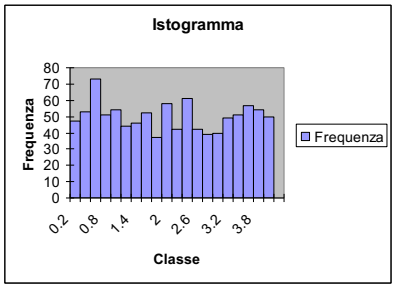


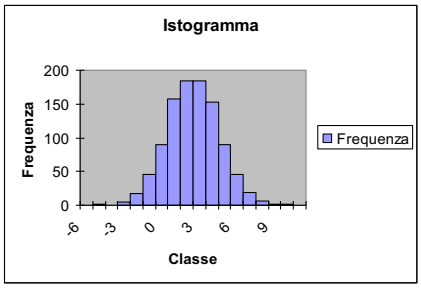
37

Media e varianza:

Media uguale

Deviazione Standard Diversa





Media=2
Varianza=1.33

Media=2
Varianza=4

38

dispersione di una distribuzione

Numero di ore di sonno	frequenza Maschi	Femmine
1	1	3
2	3	6
3	3	7
4	7	8
5	11	5
6	8	3
7	4	1
8	2	1
9	1	1
10	-	-
11	-	1
12	-	1
13	-	1
14	-	1
15	-	1

Diamo un'occhiata alla distribuzione di frequenza delle ORE DI SONNO indotte da un sonnifero, dormite da **40 maschi** e **40 femmine**.

39

dispersione di una distribuzione

- La **misura della variabilità**, permette di descrivere in modo più completo la distribuzione di una variabile.
- Le misure di tendenza centrale: **media, mediana e moda** individuano l'elemento "centrale" della distribuzione.
- Diamo, di nuovo, un'occhiata alla distribuzione di frequenza delle **ORE DI SONNO** dei 40 soggetti.
  - ✓La media è di **5 ore** ma uno sguardo alla tabella mostra che un buon numero di pazienti sono molto diversi tra loro.
  - ✓Alcuni presentano un periodo di sonno **più breve** ed altri **più lungo della media**.
- La media **non dice** in che misura i dati siano dispersi attorno al valore centrale.

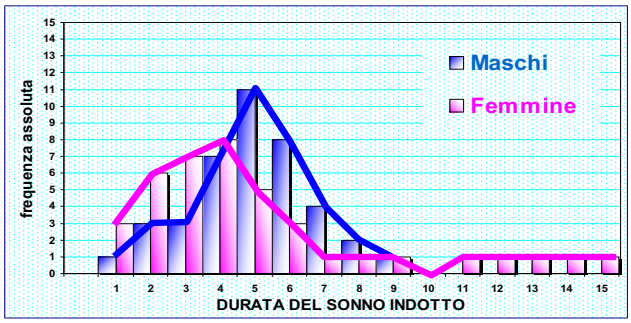
40

dispersione di una distribuzione

Il numero medio di "letture" risulta di 5 ore in entrambe i sessi

Uguale *durata del sonno indotto* ?

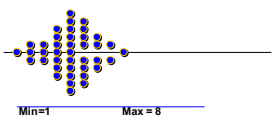
Per facilitare i confronti riportiamo i dati in grafico.

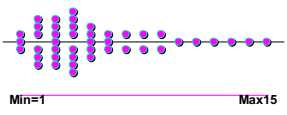


41

L'intervallo di variazione

- Mentre **in media** le femmine presentano un durata del sonno uguale ai maschi, alcune di loro hanno un durata del sonno ancora superiore ai tempi più elevati dei maschi.
- Quindi le medie non sono insufficienti: per completare il quadro occorrono alcune misure di variabilità.
- L'**intervallo di variazione** o range consiste semplicemente nella differenza tra il valore massimo e il valore minimo della distribuzione.





42

L'intervallo di variazione

Esempio:

Gli insiemi di valori di VES

{A}: { 8, 5, 7, 6, 35, 5, 4}
{B}: { 11, 8, 10, 9, 17, 8, 7}

hanno la stessa media ($\bar{x}$=10),

ma in {A} i valori sono più dispersi che in {B}:

in {A} i valori sono inclusi tra 4 e 35

in {B} i valori sono inclusi tra 7 e 17

La differenza tra il massimo e il minimo valore di un insieme di dati è detto **intervallo di variazione** (o **range**).

il **range** di {A} è $R_A = 35 - 4 = 31$

il **range** di {B} è $R_B = 17 - 7 = 10$

Il **range** è il più ***intuitivo*** fra gli indici di dispersione, ha però il difetto di basarsi solo sui due valori estremi, nei quali si manifesta maggiormente la variabilità di campionamento e l'errore di misura.

43

La devianza

Gli indici di dispersione di più largo uso sono basati sugli **scarti dalla media**: per un campione di dimensione n, $\{x_1, x_2, \ldots x_n\}$, sono così definiti

Devianza: $D = \sum (x_i - \bar{x})^2$

Varianza campionaria: $s^2 = \frac{D}{n-1}$

Deviazione standard: $s = \sqrt{s^2}$

Coefficiente di variazione: $CV\% = 100 \times \frac{s}{\bar{x}}$

La **devianza** è la somma dei quadrati degli scarti tra ogni elemento del campione (x_i) e la media campionaria ($\bar{x}$).

44

formule di calcolo della devianza

devianza per dati singoli

$$D=\sum_{i=1}^{n}(x_i-\bar{x})^2$$

$$= \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$$

devianza per dati raggruppati in classi

$$D = \sum (x_i-\bar{x})^2 f(x_i)$$

$$= \sum x_i^2 f(x_i) - \frac{\left(\sum x_i f(x_i)\right)^2}{\sum f(x_i)}$$

45

calcolo degli indici di dispersione

- Nell'***esempio*** dei due insiemi di valori di VES si ha:

{A}: { 8, 5, 7, 6, 35, 5, 4}

$D = 8^2+5^2+\ldots 4^2 - (8+5+\ldots 4)^2/7 = 1440-700=740$

$s^2 = 740/6 = 123.33$ $s = \sqrt{123.3} = 11.1$ i ={-1.1,21.1}

CV%= 100 (11.1/10) = 111%

{B}: { 11, 8, 10, 9, 17, 8, 7}

$D = 11^2+8^2+\ldots 7^2-(11+8+\ldots+7)^2/7 = 768-700 = 68$

$s^2 = 68 / 6=11.33$ $s=\sqrt{11.33} = 3.4$ i ={6.6, 13.4}

CV% = 100 (3.4/10) = 34%

In {A} l'intervallo ± s include anche valori negativi di VES, che ovviamente non sono possibili. L'uso di s per esprimere la dispersione dovrebbe essere quindi limitato alle **distribuzioni simmetriche** (o quasi).

46

calcolo della devianza (dati in classi)

limiti di classe	x_i	$f(x_j)$	$x_i f(x_j)$
44.25- 45.75	45.0	2	90.0
45.75- 47.25	46.5	5	232.5
47.25- 48.75	48.0	7	336.0
48.75- 50.25	49.5	14	693.0
50.25- 51.75	51.0	16	816.0
51.75- 53.25	52.5	9	472.5
53.25- 54.75	54.0	5	270.0
54.75- 56.25	55.5	1	55.5
56.25- 57.75	57.0	1	57.0
		60	3022.5

47

calcolo della devianza (dati in classi)_5di5

Nell'**esempio** della lunghezza dei neonati:

x_i	$f(x_i)$	$x_i f(x_i)$	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 f(x_i)$	x_i^2	$x_i^2 f(x_i)$
45.0	2	90.0	-5.375	28.891	57.781	2025.00	4050.00
46.5	5	232.5	-3.875	15.016	75.078	2162.25	10811.25
48.0	7	336.0	-2.375	5.641	39.484	2304.00	16128.00
49.5	14	693.0	-0.875	0.766	10.719	2450.25	34303.50
51.0	16	816.0	0.625	0.391	6.250	2601.00	41616.00
52.5	9	472.5	2.125	4.516	40.641	2756.25	24806.25
54.0	5	270.0	3.625	13.141	65.703	2916.00	14580.00
55.5	1	55.5	5.125	26.266	26.266	3080.25	3080.25
57.0	1	57.0	6.625	43.890	43.890	3249.00	3249.00
	60	3022.5			365.812		152624.25

media= 3022.5 /60= 50.375

D = $(45.0-50.375)^2 \cdot 2 + (46.5-50.375)^2 \cdot 5 + \ldots + (57.0-50.375)^2 \cdot 1 = 365.812$

D = 152624.25 - $(3022.5)^2/60$ = 152624.25 - 152258.44 = 365.813

Var= 365.812/59 =6.2

Deviazione standard = 2.49

48

calcolo della varianza (dati in classi)

x_i	$f(x_i)$	x_i^2	$x_i \cdot f(x_i)$	$x_i^2 \cdot f(x_i)$	$(x_i - \bar{x})$	$(x_i - \bar{x})^2 \cdot f(x_i)$
1	4	1	4	4	-4	64
2	9	4	18	36	-3	81
3	10	9	30	90	-2	40
4	15	16	60	240	-1	15
5	16	25	80	400	0	0
6	11	36	66	396	1	11
7	5	49	35	245	2	20
8	3	64	24	192	3	27
9	2	81	18	162	4	32
10	0	100	0	0	5	0
11	1	121	11	121	6	36
12	1	144	12	144	7	49
13	1	169	13	169	8	64
14	1	196	14	196	9	81
15	1	225	15	225	10	100
Σ	80		400	2620		620

Devianza= 620 ; Varianza=Devianza/(N-1)= 620/79 = 41.33

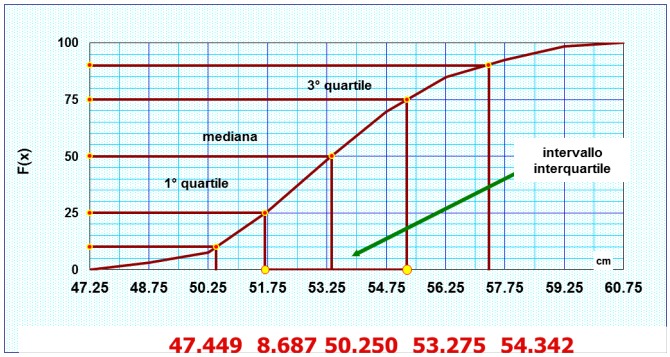
Torniamo all'esempio delle ORE DI SONNO

Deviazione standard=6.429

49

l'intervallo interquartile

Un indice di dispersione di uso comune è l'**intervallo interquartile**, dato dalla **differenza tra 3° e 1° quartile** (cioè tra 75° e 25° centile): tale intervallo contiene la metà dei valori inclusi nel campione, indipendentemente dalla forma della distribuzione della variabile.



Indici di dispersione:

$x_{max} - x_{min}$	Range (intrevallo di variazione)
$\frac{1}{n}\sum_1^n \lvert x_i - \mu \rvert$	Scarto medio assoluto
$\frac{1}{n}\sum_1^n (x_i - \mu)^2$	Media dei quadrati degli scarti
$\frac{1}{n-1}\sum_1^n (x_i - \bar{x})^2$	Varianza campionaria
$\sqrt{\frac{1}{n-1}\sum_1^n (x_i - \bar{x})^2}$	Deviazione standard campionaria

p_esimo quantile: si considera np per [0 ≤ p ≤1]
Se np non è intero, considero k l'intero successivo e il p_esimo quantile è x_k
Se np = k è intero, il p_esimo quantile è $(x_k + x_{k+1})/2$

Q_1=primo quartile =25° percentile
Q_2=secondo quartile =50° percentile =mediana
Q_3=terzo quartile =75° percentile

51

Principali indici statistici

I grafici finora analizzati ci danno informazioni qualitative; possiamo quantificarle ricorrendo ai seguenti indici.
Siano $x_1, x_2, \ldots, x_n$ n osservazioni numeriche

- di posizione
  - MODA
  - MEDIANA
  - MEDIA
- INDICI
  - di dispersione
    - SCARTO QUADRATICO MEDIO
    - VARIANZA
    - RANGE
  - di forma
    - ASIMMETRIA (SKEWNESS)
    - CURTOSI (KURTOSIS)

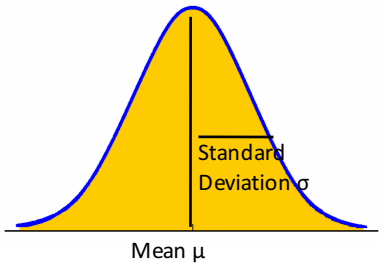
52

La distribuzione normale

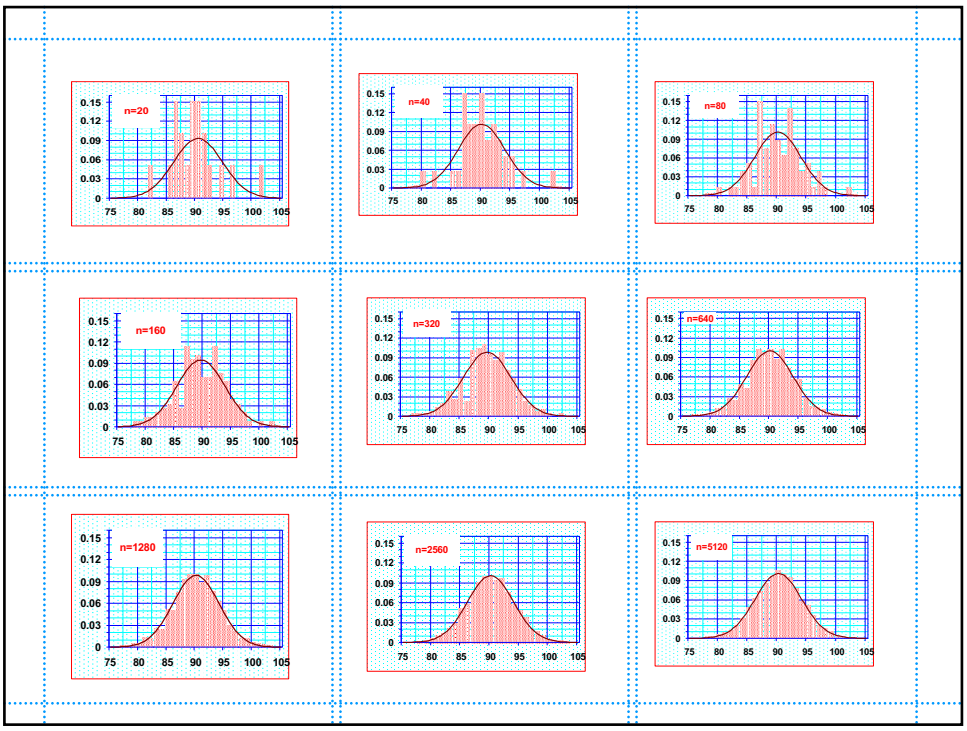
Johann Carl Friedrich Gauss
(1777-1855)

LA FORMA DELLA DISTRIBUZIONE DEGLI ERRORI DI MISURA

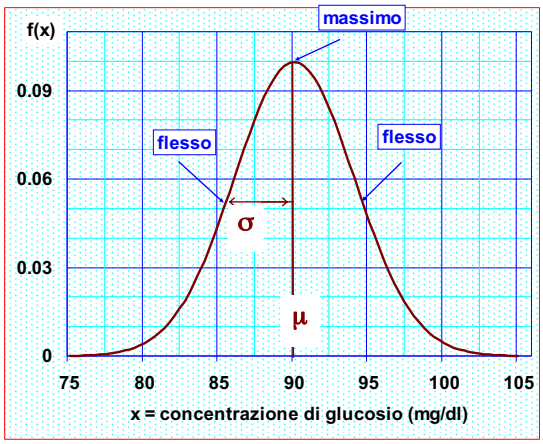
All'aumentare del numero di misure, i valori tendono ad accentrarsi attorno alla loro media e l'istogramma assume una forma ***a campana*** sempre più regolare, che può essere approssimata con una funzione reale nota come **funzione di gauss**/ **funzione normale**.



Johann Carl Friedrich Gauss
(1777-1855)



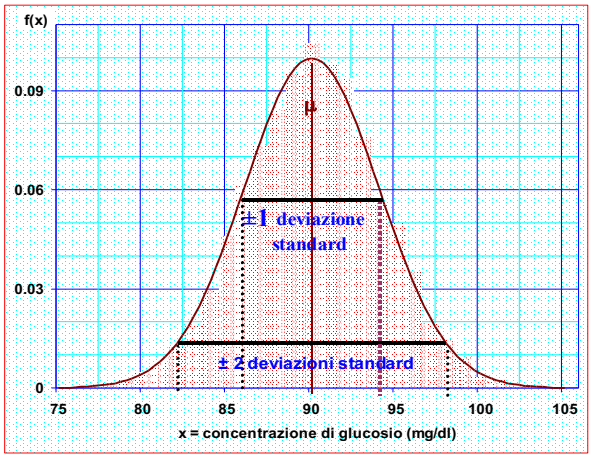
La funzione di Gauss (3)



$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

dove: **?** è la deviazione standard della totalità delle misure;
μ è la media della totalità delle misure;
e* = *base dei *logaritmi naturali* ($e = 2.71828\ldots$).
π è il rapporto tra circonferenza e diametro ($\pi = 3.14159\ldots$);

La funzione di Gauss (2)



La funzione di Gauss (1)

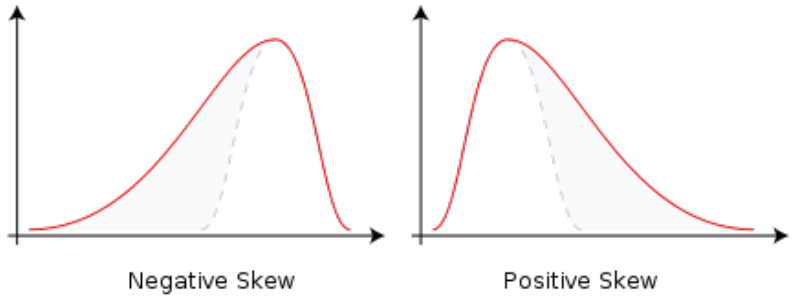
- Gli errori casuali di misura. considerati nel loro complesso, mostrano un comportamento tipico che può essere così descritto:

- Gli ***errori piccoli*** sono più frequenti di quelli ***grandi***;
- Gli errori di ***segno negativo*** tendono a manifestarsi con la stessa frequenza di quelli con segno positivo;
- All'aumentare del numero delle misure si ha che circa 2/3 dei valori tendono ad essere inclusi nell'intervallo **media +/- 1 deviazione standard**
- Il **95%** dei valori tende ad essere incluso nell'intervallo **media +/- 2 deviazioni standard**

Quali sono I migliori descrittori statistici per un campione?

Dati estratti da pdf Gaussiana: Media +/- Deviazione Std

Skewness

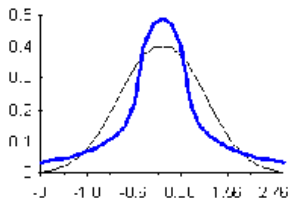


60

CURTOSI: leptocurtica

Curtosi: distribuzione leptocurtica

Distribuzione leptocurtica: una frequenza maggiore dei valori estremi e dei valori centrali.

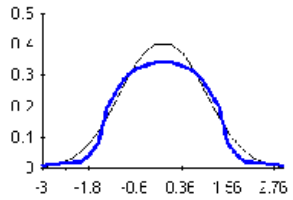


CURTOSI:

distribuzione platicurtica

Curtosi

Distribuzione platicurtica: frequenza minore dei valori estremi e dei valori centrali.



Indici di forma

Skewness

INDICE DI ASIMMETRIA

$$\gamma_1 = \frac{m_3}{m_2^{3/2}} = \frac{\sum (x_i - \mu)^3}{n\sigma^3}$$

>0 coda a destra
<0 coda a sinistra
=0 simmetrica

Per la distribuzione gaussiana γ=0

Kurtosis

$$g_2 = \frac{m_4}{m_2^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2}$$

Misura quanto la distribuzione è appuntita
>3 poco appuntita
=3 caso della distribuzione normale
<3 molto appuntita

Per la distribuzione gaussiana g_2=3

63

Coefficienti di skewness di Pearson

Karl Pearson ha suggerito i calcoli più semplici come una misura di asimmetria:

La modalità di asimmetria di Pearson, definito da

- (media - Moda) / deviazione standard,

Asimmetria primo coefficiente di Pearson, definita da

- 3 (media - moda) / deviazione standard,

Asimmetria secondo coefficiente di Pearson, definito da

- 3 (media - mediana) / deviazione standard.

64

Quali sono i migliori descrittori statistici per un campione?

Dati estratti da pdf Gaussiana: Media +/- Deviazione Std

Dati estratti da pdf Non-Gaussiana: [Mediana +/- Int. Interq, 3° misura]
(range, skewness, kurtosis, etc)

Applicazioni

- La simmetria ha benefici in molti settori. In molti modelli è semplicistico supporre che i dati abbiano una distribuzione *[normale]* simmetrica intorno alla media.
- La distribuzione normale ha una asimmetria di zero. Ma in realtà, spesso i punti dati non sono perfettamente simmetrici.
- La comprensione dell'asimmetria della serie di dati reali indica che le deviazioni dalla media stanno più nel verso positivo o più nel verso negativo.
- Il test K^2 (D'Agostino) è un Goodness-of-fit test di normalità basato sulla asimmetria e curtosi campionaria.

66

Indici: Schema riassuntivo

di posizione

- **media:** $\bar{x} = \frac{\sum_i x_i}{N}$
- **moda:** punto di max della distribuzione
- **mediana:** valore sotto al quale cadono la metà dei valori campionari. Si dispongono i dati in ordine crescente e si prende quello che occupa la posizione centrale (N dispari) o la media dei 2 valori in posizione centrale (N pari)

di dispersione

- **varianza** $s^2 = \frac{\sum_i \left(x_i - \bar{x}\right)^2}{N-1}$
- **deviazione standard** s
- **range** $R = x_{max} - x_{min}$

di di forma

- **skewness (coeff. di asimmetria)** $\frac{\sum_i \left(\frac{x_i - \bar{x}}{\sigma}\right)^3}{N}$

>0 coda a ds
<0 coda a sin
=0 simmetrica

- **curtosi:** misura quanto la distribuzione è appuntita $\frac{\sum_i \left(\frac{x_i - \bar{x}}{\sigma}\right)^4}{N}$

>3 poco appuntita <3 molto appuntita

67